Manuscript Title

This manuscript (permalink) was automatically generated from greenelab/generalization-manuscript@7b5e195 on July 12, 2023.

Authors

- **John Doe**
  XXXX-XXXX-XXXX-XXXX · johndoe · johndoe · @johndoe@mastodon.social
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Jane Roe** ✉
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

✉ — Correspondence possible via GitHub Issues or email to Jane Roe <jane.roe@whatever.edu>.

Abstract

Results

“Best” and “smallest good” model selection strategies perform comparably

To address the question of whether more parsimonious models tend to generalize better or not, we designed two model selection schemes and compared them for the TCGA to CCLE and CCLE to TCGA mutation prediction problems (Figure 1A). The “best” model selection scheme chooses the top-performing model/LASSO parameter on the holdout dataset from the same source as the training data and applies it to the test data from the other data source. The intention of the “smallest good” model selection scheme is to balance parsimony with reasonable performance on the holdout data, since simply selecting the smallest possible model (generally, the dummy regressor/mean predictor) is not likely to generalize well. To accomplish this, we first identify the top 25% of well-performing models on the holdout dataset; then, from this subset of models, we choose the smallest (i.e., highest LASSO parameter) to apply to the test data. In both cases, we exclusively use the holdout data to select a model and only apply the model to out-of-dataset samples to evaluate generalization performance *after* model selection.

For TCGA to CCLE generalization, 27/71 genes (38.0%) had better performance for the “best” model, and 17/71 genes (23.9%) had better generalization performance with the “smallest good” model. The other 27 genes had the same “best” and “smallest good” model (in other words, the “smallest good” model was also the best-performing overall, and the difference was 0) (Figure 1B). For CCLE to TCGA generalization, 23/70 genes (32.9%) had better performance for the “best” model and 18/70 (25.7%) for the “smallest good,” with the other 29 having the same model fulfill both criteria (Figure 1C). Overall, these results do not support the hypothesis that the most parsimonious model generalizes the best: for both generalization problems there are slightly more genes where the best-performing model on the holdout dataset is also the best-performing on the test set, although there are some genes where the “smallest good” approach works well.

We examined genes that fell into either category for TCGA to CCLE generalization (dotted lines on Figure 1B). For *NF1*, the “best” model outperforms the “smallest good” model (Figure 1D). Comparing holdout (orange) and cross-dataset (green) performance, both generally follow a similar trend, with the cross-dataset performance peaking when the holdout performance peaks at a regularization parameter of $\alpha = 0.00316$. *PIK3CA* is an example of the opposite, a gene where the “smallest good” model tends to outperform the “best” model (Figure 1E). In this case, the peak for the cross-dataset performance occurs at a higher level of regularization (further left on the x-axis), at $\alpha = 0.01$, than the peak for the holdout performance, at $\alpha = 0.0316$. This suggests that a *PIK3CA* mutation status classifier that is more parsimonious, but that has slightly worse performance, does tend to generalize better across datasets to CCLE.

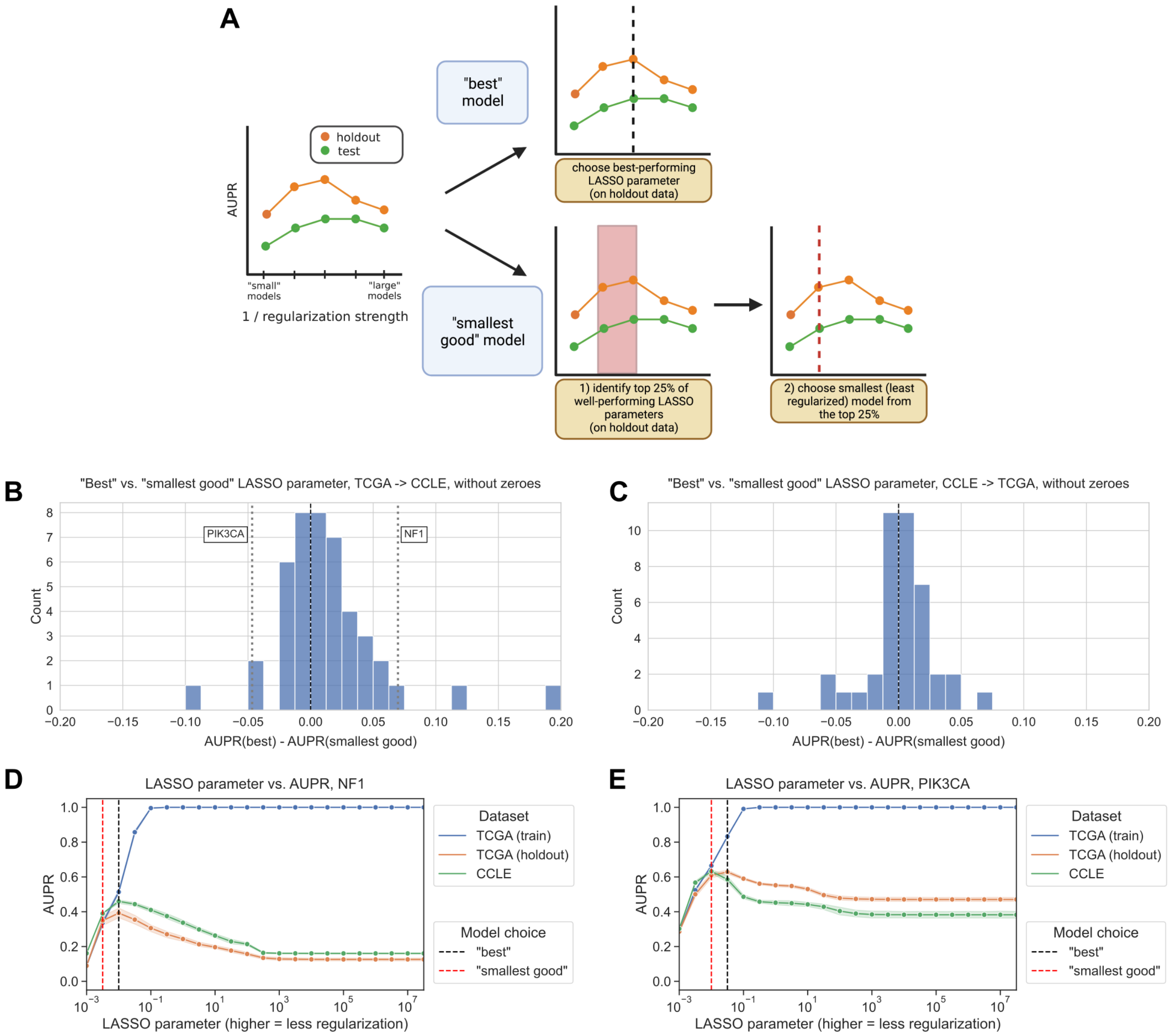

Figure 1: **A.** Schematic of "best" vs. "smallest good" model comparison experiments. **B.** Distribution of performance comparisons between "best" and "smallest good" model selection strategies, for TCGA -> CCLE generalization. Positive x-axis values indicate better performance for the "best" model, negative values indicate better performance for the "smallest good" model. **C.** Distribution of performance comparisons between "best" and "smallest good" model selection strategies, for CCLE -> TCGA generalization. **D.** *NF1* mutation status prediction performance generalizing from TCGA (holdout, orange), to CCLE (green), with "best" and "smallest good" models labeled. **E.** *PIK3CA* mutation status prediction performance generalizing from TCGA (holdout, orange), to CCLE (green), with "best" and "smallest good" models labeled.

References